PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G06F 15/00 — A1

(11) International Publication Number: **WO 00/28429**

(43) International Publication Date: 18 May 2000 (18.05.00)

(21) International Application Number: PCT/US99/25922

(22) International Filing Date: 5 November 1999 (05.11.99)

(30) Priority Data:
60/107,474 6 November 1998 (06.11.98) US

(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application
US 60/107,474 (CIP)
Filed on 6 November 1998 (06.11.98)

(71) Applicant *(for all designated States except US)*: GLAXO GROUP LIMITED [GB/GB]; Glaxo Wellcome House, Berkeley Avenue, Greenford, Middlesex UB6 0NN (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: CHEN, Xin [CN/US]; Computer Assisted Drug Discovery, R.W. Johnson Pharmaceuticals Research Institute, 1000, Route 202, P.O. Box 300, Raritan, NJ 08869–0602 (US). RUSINKO, Andrew, III [US/US]; 2502 Mandy Way, Arlington, TX 76017 (US). YOUNG, S., Stanley [US/US]; Glaxo Wellcome Inc., Five Moore Drive, P.O. Box 13398, Research Triangle Park, NC 27709 (US).

(74) Agents: LEVY, David, J.; Glaxo Wellcome Inc., Five Moore Drive, P.O. Box 13398, Research Triangle Park, NC 27709–3398 (US) et al.

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: MEANS AND METHOD FOR RECURSIVE PARTITIONING ANALYSIS OF LARGE STRUCTURE-ACTIVITY DATA SETS

(57) Abstract

The identification of three–dimensional pharmacophores from large, heterogeneous data sets is achieved by combining a conformational and correspondence search procedure. More specifically, each flexible ring in a compound is isolated by cutting off one or more side chains while keeping the side chain neighbors nearest the flexible ring. Every flexible corner of the isolated ring is perturbed relative to an average ring plane. The chains that were cut off are reconnected and torsional ranges for each rotatable bond is calculated. Responsive to the identification of points within the calculated torsial ranges, the identified conformations are uniformly sampled and stored. The conformation generation process is integrated into a correspondence search procedure to maximize both the efficiency and speed.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

MEANS AND METHOD FOR RECURSIVE PARTITIONING ANALYSIS OF LARGE STRUCTURE-ACTIVITY DATA SETS

RELATED APPLICATION

The subject matter of this application is related to the subject matter of the commonly owned application PCT Serial Number WO98/47087, entitled "Statistical Deconvoluting of Mixtures" filed on April 17, 1998, the contents of which are incorporated by reference as if fully disclosed herein.

BACKGROUND

The recent progress of combinatorial chemistry and high throughput screening techniques has brought a revolution in the drug discovery processes in the pharmaceutical industry. It is now feasible to obtain biological activity data for thousands to hundreds of thousands of chemical compounds in a short period of time, leading to the tremendous increase of the quantity of data for the drug discovery cycle. However, analysis and utilization these data sources and/or conversion into a useful formats in a timely fashion is still a challenge for chemoinformatics. Among the many relevant problems, one is the automated pharmacophore identification for large, heterogeneous chemical data sets.

A pharmacophore, usually defined as the key chemical features and the spatial relationships between them (configurations) associated with the biological activities of chemical compounds, is one of the most important concepts in medicinal chemistry and plays a critical role in the drug discovery process. Pharmacophore models can help medicinal chemists gain an insight on the key ligand-receptor interactions that are responsible for the biological activities, even when the receptor structure has not been determined. The models can be used as the search queries for pharmacophore search thereby assisting in the discovery of lead compounds. The models can also be used as the

initial step of 3D QSAR analysis by grouping the compounds that follow the same binding mode and indicating the possible 3D alignment rules.

However, pharmacophore identification remains a process heavily dependent on medicinal chemists' experience and intuition. In most of the reported work on pharmacophore search, the search queries were taken from the literature or the crystal structures of receptor-ligand complexes. Over the past decade, several efforts, called automated pharmacophore identification/mapping/recognition, have been tried to involve more contributions from computational science into the pharmacophore identification process.

Some of the algorithms and programs that have been specifically designed for this purpose include: active analogue approach, ensemble distance geometry, DISCO, Catalyst/Hypo, HipHop, Apex-3D, DANTE, and using ILP (Inductive Logic Programming) system Progol. However, each of these programs or algorithms suffer one or more of the following limitations: a) they are inherently limited to small data sets, which typically contain less than 50 compounds, since none of the programs were originally designed for large, heterogeneous chemical data sets; b) the programs only utilize the structural information provided by a small number of active compounds; c) most of these algorithms can not handle the situation of multiple binding modes, which are expected in large chemical data sets. Preferably, any software designed to remedy these shortcomings can also complete this process in a reasonable amount of time enabling quicker identification of pharmacophore models.

What is needed is a system and method for identifying three-dimensional pharmacophores from large, heterogeneous data sets, while utilizing structure and activity information from a large array of compounds while completing the process in a reasonable amount of time.

SUMMARY OF INVENTION

The present invention, whose software embodiment is referred to as SCAMPI (Statistical Classification of Activities of Molecules for Pharmacophore Identification), identifies three-

dimensional pharmacophores from large, heterogeneous data sets by combining fast conformation generation with recursive partitioning. This is an extension on the SCAM (Statistical Classification of Activities of Molecules) software system. The pharmacophore identification process runs recursively and the conformation spaces are re-sampled under the constraints of the evolving pharmacophore model. The present invention derives pharmacophore models from data sets up to 2000 compounds, with thousands of conformations generated for each compound. With the improvement in efficiency generated by the present invention, this process can be completed in less than one day of computational time. Thus, the present invention enables fast computation of pharmacophores from large, structurally heterogeneous data sets. The identified pharmacophores can then be used for drug design, as input to computational chemistry methods like 3D QSAR and the mathematical/in silica screening of large 3D databases of real or virtual compounds.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 provides an example of a band contact and the release procedure;

Figure 2 is a flowchart illustrating the conformation search method of the present invention;

Figure 3 illustrates the conformation generation procedure for a pseudo-compound using the present invention;

Figure 4 is a flowchart of the general design of the software implementation of the present invention, referred to as SCAMPI, wherein the conformational and correspondence searches are combined;

Figure 5 illustrates a tree generated by SCAMPI for the MAO data set;

Figure 6 illustrates the pharmacophores for the MAO data set;

Figure 7 illustrates a tree generated by SCAMPI for the ACE data set; and

Figure 8 illustrates the pharmacophores for the ACE data set.

Further illustrations are also included that provide an additional overview of the present invention.

DESCRIPTION

Before going into the details of the software embodiment of the present invention, information concerning the search queries enabled by the present system and method is helpful. Generalized chemical features that are important in selective binding are usually used in the pharmacophore search queries. In the preferred embodiment, the definition system for such chemical features includes the following features as default: negative charge centers, positive charge centers, hydrogen bond acceptors, hydrogen bond donors, aromatic ring centers, hydrophobic centers, triple bond center, plus some explicit atom types for heteroatoms: N, O, S, P, F, and halogen (except F). To attain both generality and efficiency, these chemical features are determined in two ways: *substructure search* using a fragment library as queries for the chemical features defined by multiple atoms, like guanidine, etc., and general rule-based searches for the chemical features defined by a single atom and its closest neighbors. Additional definitions can be added into the fragment library to include new chemical features.

The pharmacophore search and identification process traditionally involves searching two spaces: *the conformational space* that represents all the reasonable 3D structures for each individual compound and *the correspondence space* that indicates all the common chemical feature matchings for a class of compounds.

(1) Conformational search.

As herein described, all the structure data for the conformational search is automatically generated by a topology analysis. During a topological analysis, the topological structure of each compound is decomposed into small units, which is the maximal subsets of atoms whose inter-atomic distances are invariant with respect to the torsional rotation of rotatable bonds. There are two types of units defined in SCAMPI: rigid groups and flexible rings. Rigid groups have no additional internal conformational freedom degrees, for example the aromatic ring, methylene group, and methyl group, while flexible rings can change their own conformations by ring flipping, such as the cyclohexane ring. The flippable corners on each flexible ring are also identified as the non-fusion ring atoms.

Once identified, these units are assigned sequential numbers that determine the order in which they will be assembled in the conformation build-up process.

The conformational search implemented by the present system and method provides a number of advantages over the prior art methods. Traditionally, conformational search methods have tried to use a small set (usually tens) of conformers to completely cover the whole conformational space of each compound. These methods are hopeful at best. The representative conformers were usually generated by some conformation search method followed by clustering analysis. Alternatively, some computational methods were designed to automatically generate diverse conformers, such as the "poling" method. Force field calculation and energy minimization are often used to generate low-energy conformations, typically within several kcal/mol of the global minimum. A Tripos-like force field was set up for the molecular mechanism calculations in SCAMPI. It includes terms for bond, angle, torsion, plane and van der Waals interactions, except that the electrostatic interaction term is excluded to simplify and speed the calculations. All the data structures necessary for force field calculations are also automatically generated by topology analysis. The drawbacks of this strategy are obvious. Tens of conformers may not be enough to adequately sample the conformational space of a highly flexible compound, receptor-bound conformations may not locate in the low-energy regions, and heavy computational burdens have traditionally prohibited the use of large data sets. The present invention, however, provides a number of means for improving the operation and efficiency of the sampling process enabling the use of larger, and therefore more accurate, data sets. Each improvement will be described separately but, in conjunction with one another, they serve to provide an entirely new range of acceptable sizes of the data sets that can be used.

Referring now to Figures 2 and 3, The flow chart of the general conformational search procedure in SCAMPI is illustrated.

Step 1: If there are any flexible rings in the compound, each flexible ring is isolated by cutting off its side chain(s) but keeping its nearest side-chain neighbors.

In prior art methods, the conformational search of a flexible ring was transformed to conformational search of a chain by breaking one of the ring bonds and then adding a tight distance constraint between the two terminal atoms of that broken bond. Obviously, this kind of distance constraint needs to be a very tight constraint, and it can be expected that considerable computational effort is required for the successful ring closure. However, conformational search can be implemented in two different coordinate systems: Cartesian coordinate system where the Cartesian coordinates of each atom are directly perturbed, and internal coordinate system where the torsion angle of each rotatable bond is directly modified. Cartesian coordinates are more suitable for the conformational search of flexible rings, while internal coordinates are a natural choice for the non-ring parts of the structure. For this reason, the conformational search in SCAMPI is implemented in both Cartesian and internal coordinate systems, for searching conformations of flexible rings and chains, respectively.

Step 2: Each flexible corner of the isolated ring is randomly perturbed relative to the average ring plane. Several steps (20-100 steps) of energy minimization are then used to optimize the ring structure to a reasonable geometry.

Step 3: All the side-chains are re-connected to the ring to form a complete compound. Another several steps (20-100 steps) of energy minimization are used to optimize the whole compounds structure to release the possible bad van der Waals contacts between side chains and rings.

Step 4: Now, treating the flexible rings as rigid, the "differential distance equation" algorithm is sequentially used to find acceptable torsional ranges for each rotatable bond. After the acceptable torsional ranges of a rotatable bond are determined by the "differential distance equation" algorithm, the sampling points within such acceptable torsional ranges are randomly and uniformly picked up. The random search is performed in order to obtain a uniform sampling of the whole conformational space. More easily accessible conformations

are more likely to be shared by the active compounds at the binding site, and therefore should give stronger signal and have more chance to be selected by the statistical test.

Description:

The conformational search method applied in SCAMPI utilizes the "differential distance equation" to generate conformational structures that do not contain any bad van der Waals bumps between atoms. "Differential distance equation" algorithm is a look-ahead algorithm. It can determine the acceptable torsional ranges of a rotatable bond, which can lead to the partial conformational structure satisfying various distance constraints, such as the van der Waals distance constraints (and, as will become apparent, importantly the pharmacophore distance constraints) before it actually rotates that bond. Thereby it is much more efficient than the usual trial-and-error algorithms, where most of the computational efforts are spent to find the acceptable torsional angles.

Previously, the acceptable torsional ranges of a rotatable bond determined by the intersection of acceptable torsional ranges of all the atom pairs that contain atoms on the both sides of that bond. Thus, if there are M atoms on one side of a rotatable bond and N atoms on the other side, we need repeat the calculation for about MxN times before we can get the final acceptable torsional ranges. In most cases, however, the final acceptable torsional ranges can be determined by only a small number of atom pairs (usually far less than MxN). Thus, the acceptable torsional ranges of many atom pairs are $[-\pi, \pi]$, and therefore they need not be considered further in the intersection step. This results from the fact that the minimum accessible distances between the two atoms of such atom pairs are larger than their van der Waals distance constraints, so that it is not possible to have van der Waals contact between these atom pairs no matter how the bond rotates.

Consequently, the present invention uses the minimum-accessible-distance calculation to examine all the atom pairs. This calculation is much faster than the calculation of the acceptable torsional ranges by the "differential distance equation". Only those atom pairs whose minimum accessible distances are less than their van der Waals distance constraints will be further considered by the "differential distance equation" algorithm. After the acceptable torsional ranges of a rotatable bond are determined by the "differential distance

equation" algorithm, the sampling points within such acceptable torsional ranges are randomly and uniformly picked up. The inclusion of such a minimum-accessible-distance calculation as a filter significantly increases the computational speed.

Step 4: (continued) If the acceptable torsional ranges are empty, the "release procedure" will be tried to relieve the existing bad contacts. If this release attempt fails, the whole build-up process will be started again from the first rotatable bond. To save the computational cost, the ring conformational search is completed only periodically, only after the chain conformational search has been successfully computed several times (default 5).

Description: One problem with the "differential distance equation" algorithm is that is can only look one-step ahead, which means that the conformation build-up process may become stuck with a bad conformation. As illustrated in Figure 1, the process is obviously stuck at the 5th rotatable bond. No matter how you rotate the 5th bond, there is always bad contact between the two terminal units. Prior art methods will return to sample the next sampling point of the 4th rotatable bond, to see if this bad contact can be released. If all the sampling points of the 4th rotatable bond have been tried and the bad contact still exists, the algorithm will go back to the 3rd rotatable bond. This backtracking strategy continues until the bad contact is released. At times, however, the bad contacts, as shown in Figure 1, could be released by rotating some rotatable bond between these two conflicting units. This release process can also be realized by using the "differential distance equation" algorithm with re-grouping the units, as illustrated to the 3rd bond in Figure 1. When this release strategy fails to resolve bad contact problems, the present invention will restart the conformational build-up procedure from the 1st rotatable bond. In most cases, this release strategy works very well and no further rebuild-up procedure is needed.

(2) Correspondence search.

The earliest pharmacophore identification strategies, like active analogue approach and ensemble distance geometry, avoided the correspondence search by requiring the user to

identify the correspondence relationships of the pharmacophoric features among different active compounds. More recently, most of the strategies and programs depended on some pair-wise comparison algorithm to determine the common chemical features and configurations in all of the active compounds. (Start with one pair of compounds and identify corresponding features.) This kind of strategy is computationally intensive and inherently limits itself to small data sets, since with n compounds at least n-1 times of pair-wise comparisons are needed, (assuming that the most active compound is used as one compound in the pair).

The correspondence search method applied in SCAMPI is based on our previous recursive partition work using FIRM and SCAM programs (see included reference for further details). The method utilizes the information on the biological activities of all the compounds in the training set, and identifies the pharmacophore(s) by detecting the structural features that are most statistically significantly correlated with the biological activities. An easily interpreted dendrogram or tree diagram is also generated, in which the statistically best structural descriptors are used to split the large data set into smaller and more homogeneous subsets. The advantages of recursive partition strategy include: a) It is inherently fast when compared grouping with many other methods for grouping compounds; It overcomes the difficulties of handling nonlinear relationships and strong interactions in large SAR data sets; and c) It can detect the multiple mechanisms by separating the chemical compounds with different mechanisms into the different arms and terminal nodes of the dendrogram.

The Student's t-test is used to recursively partition the whole data set into smaller and more homogeneous subsets, until each subset can not be split any longer. If the compounds are scored active/inactive rather than a continuous potency, then a chi-square test can be used rather than a t-test. For other methods of recursive partitioning, such as CART and C4.5, using the t-test (or chi-square test) is the preferred method to split a node. A string of "binary" descriptors is generated at first to describe each compound, which indicate the presence or absence of a series of structural descriptors in a compound. Then, each one of all the structural descriptors is checked sequentially and the data set is split into

two subsets according to whether or not that descriptor is in the structure. The Student's t-test is computed according to the following formula:

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{1}{M} + \frac{1}{N}} \cdot \sqrt{\frac{SSX + SSY}{M + N - 2}}}$$

where

$$SSX = \sum_{i=1}^{M} \left(X_i - \overline{X}\right)^2 \qquad \overline{X} = \sum_{i=1}^{M} X_i / M$$

$$SSY = \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2 \qquad \overline{Y} = \sum_{i=1}^{N} Y_i / N$$

$X_1, X_2, \ldots X_M$ are the activities in the first subset, and $Y_1, Y_2, \ldots$ are the activities in the second subset. M and N are respectively the numbers of compounds in these two subsets. The structural descriptor that gives the largest t-value is chosen as the best descriptor for the split, if its corresponding Bonferroni adjusted p-value is smaller than some pre-set termination criterion (default 0.01). The Bonferroni adjustment multiplies the raw Student t-test's p-value by the number of variables under consideration, taking into account the number of statistical tests in order to avoid the increased probability of a false positive split.

Since multiple conformations have been generated for each compound, the method assigns the absence of a particular descriptor to a compound if none of its conformations contains that descriptor, otherwise, if any of its conformations contains that particular descriptor, the method assign the presence of it to that compound. This is akin to a Boolean OR operation on all the *conformation* "binary" descriptor strings of a compound, forming the *compound* "binary" descriptor string, which takes the form of a bit, 0/1, string.

The pharmacophore model construction procedure of the present invention proceeds in the following fashion: a two-point pharmacophore is searched first and then the new pharmacophore point is searched and added if it is found. This process continues adding a single point at a time until no more statistically significant pharmacophore points can be found. Correspondingly, two-point structural descriptors are generated for each compound at first, which are composed of two chemical features and the "binned" distance between

them. After the most significant two-point descriptor has been determined, new three-point structural descriptors are generated for each compound. These three-point structural descriptors retain the features and distance in the most significant two-point descriptor that has been found, and therefore are actually composed of only the third chemical feature and its two **"binned"** distances to each one of the former two chemical features. Following the same rule, the newer descriptors are composed of a newer chemical feature and its "binned" distances to all of the former chemical features that have been determined. This process is iterative and stops when no new pharmacophoric features can be found.

There are two kinds of possible splits at each split point: positive and negative splits. A positive split is one in which the sub-node that contains the split descriptor is the more active node, i.e., the subset of compounds containing the split descriptor is more active on average than the subset of compounds not containing that split descriptor, otherwise it is a negative split. The software implementation of the present invention uses the positive split as the default split method. Although information on the excluded volume could be determined by enabling selection of negative splits, the software implementation has the advantage of simplicity as the negative splits are often difficult to select using classic pharmacophore modeling. The default is positive splits only, but the user can ask for positive and negative splits.

(3) Combining conformational search and correspondence search.

Previous methods did not combine conformational search and correspondence search together. The present invention combines the conformational generation and correspondence search together. The sampling completeness of the descriptors is used as the criterion to terminate each round of conformational search, because further conformational searches will not add more information for the following statistical test since that particular descriptor space has been completely sampled. On the other hand, the pharmacophonic descriptors that have been identified are imposed as the additional constraints in the next-round conformational search, so that the sampling in those important conformational subspaces can be more thorough.

The general design of the software implementation of the present invention, SCAMPI, wherein the conformational and correspondence searches are combined is illustrated as the flow chart in Figure 4 and described as the following steps.

Step 1: The first-round conformational search is done for each compound without any constraints except the chemical structure itself. The search completeness is judged by monitoring the "extended" two-point descriptors, which is defined as ' (chemical feature ID)-(binned distance)-(chemical feature ID)". When there is no new "extended" two-point descriptor found for continuous n (default 100) times, it is assumed that this descriptor space has been completely sampled. All the "extended" two-point descriptors are then converted to the "standard" two-point descriptors as we have described, which has the form as "(chemical feature type)-(binned distance)-(chemical feature type)", for the following recursive partition analysis. During the conformational search process, all the generated conformations are saved in the memory space for the further use.

Step 2: Student's t-test is used to analyze the two-point descriptor space, to find the most significant two-point descriptor to split compounds into the more active and less active groups. If such a descriptor is found, the two chemical features and the distance between them in this descriptor will be treated as the first two pharmacophoric points, and the whole data set be divided into two subsets: the compounds containing this two-point descriptor and the compounds not containing it.

Step 3: The Ullman algorithm is used to search all the saved conformations of the compounds in the positive subset, using the most significant two-point descriptor as the search query. The entities of the pairs of chemical feature points, which satisfy the search query, are identified and recorded for the following constrained conformational search.

Step 4: The second-round conformational search is done for all the compounds in the positive subset. The "binned" distance in the most significant two-point descriptor is added to the pair of chemical features points, which have been identified previously by Ullman algorithm, as an additional distance constraint. Then, like in steps 1 and 2, three-point descriptors are generated during the constrained conformational search, and the search completeness is judged by monitoring "extend" three-point descriptors. Student's t-test is

used again to analyze the "standard" three-point descriptors in order to find the third most significant pharmacophoric point.

Step 5: For the compounds in the negative subset, Student's t-test is used again to analyze the former "standard" two-point descriptors in order to find a next most significant two-point descriptor.

Step 6: The above steps are repeated as shown in Figure 4, until no more significant pharmacophoric points can be found or the default maximal number of pharmacophoric points (presently, set at 5 points) has been attained.

As a further refinement on the implementation of the invention, sparse matrix techniques are utilized throughout in order to conserve memory. Lists of structures where descriptors are found, instead of lists of descriptors that are found in those structures, are stored for the recursive partitioning analysis. Hash-table search is used to insert all the descriptors into the SAR table. Furthermore, dynamic memory management is widely applied to optimize the memory utilization and also increase the computational speed since we avoid saving a large amount of conformational information on hard disk.

Examples:

(1) Monoamine oxidase (MAO) inhibitors.

Of the 1,650 compounds in the original MAO data set provided by Abbott Laboratories, CONCORD successfully converted 1,644 compounds from the 2D structures to the 3D structures. The structures and activities of these 1,644 compounds were then used as the input for SCAMPI.

With a single run of SCAMPI, a recursive partition tree was generated as shown in Figure 5. The computational time was about 1.2 CPU hours and a total of 400,033 conformations were generated in the entire process. From Figure 5, we can find two major active nodes, shaded in shadow. The two corresponding pharmacophores are illustrated in Figure 6. The first one contains an aromatic ring center, a triple bond center and a positive charge center on nitrogen. The second one contains two hydrogen bond donors on nitrogen,

with a perfectly correlated carbonyl group at the adjacent position; this group of atoms forms a hydrazide feature.

These two pharmacophore models are supported by previous experimental work. Hydrazide MAO inhibitors (e.g., compound AL16432 in Figure 6) can be hydrolyzed to acetylhydrazines that act as the non-selective, irreversible inhibitors to colvantly bind to various macromolecules including MAO. Propargylamines (e.g., compound AL19120 in Figure 6) are themselves suicide inhibitors that irreversibly inhibit MAO through covalent attachment to its flavin cofactor. Simultaneously finding features governing these two mechanisms is a clear demonstration that SCAMPI has the capability to detect multiple mechanisms of action co-existing in a large chemical data set.

(2) Angiotensin-converting enzyme (ACE) inhibitors.

The ACE data set is composed of 114 ACE inhibitors provided by Triops Inc. and 932 compounds randomly picked up from WDI (World Drug Index) database to act as negative compounds. The biological activities of 114 ACE inhibitors are expressed as continuous pIC_{50} values and the WDI compounds are arbitrarily assigned a pIC_{50} of 0. The structures and activities of these 1,046 compounds were then used as the input for SCAMPI.

With a single run of SCAMPI, a recursive partition tree was generated as shown in Figure 7. The computational time was about 8.1 CPU hours and a total of 573,798 conformations were generated in the entire process. From Figure 7, we can find two major active nodes, shaded in shadow. The two corresponding pharmacophores are illustrated in Figure 8. The first pharmacophore contains a negative charge center located on carboxylate group, an oxygen atom on carbonyl group, another negative charge center on carboxylate group, and a nitrogen atom. The second pharmacophore contains a negative charge center located on carboxylate group, an oxygen atom on carbonyl group, and a sulfur atom in thiolate group.

A comparison of the two pharmacophores in Figure 8 shows the similarity between the first three points in these two pharmacophores. They share the same geometry and two of three pharmacophonic feature types. A literature search indicates they do follow the same binding mode, by carboxylate and thiolate binding to the same zinc atom in ACE. The

commonly acceptable pharmacophore is composed of a negative charge center, an oxygen as hydrogen bond acceptor, and a zinc binding site. Because we didn't define a special chemical feature for zinc binding site, SCAMPI split the compounds following this binding mode into two different terminal nodes. As to the fourth point, a nitrogen atom, in the first pharmacophore, the statistical test indicated that it significantly contributes to the binding. This example demonstrates again that SCAMPI can quickly find the pharmacophore consistent with the known result.

In the preferred embodiment of the present invention, the system and method are implemented on a computer 900 as illustrated in figure 9. The computer comprises a central processing unit (CPU) 902 for performing the calculations of the described methods, a storage device 908 for storing data and files that can be retrieved by the processor, an input device 904 enabling user interaction with the computer, a display device 906, and dynamic memory 919 for storing one or more programs during execution, such as the program that performs the above method. Alternatively, the system and method could be implemented across a network of computers, enabling the program to be run by multiple processors at separate physical locations.

The above description is included to illustrate the operation of the preferred embodiments and is not meant to limit the scope of the invention. From the above discussion, many variations will be apparent to one skilled in the art that would yet be encompassed by the spirit and scope of the present invention.

I claim:

1. A conformational search method, wherein the steps of performing a conformational search comprise the steps of:

isolating each flexible ring in a compound by cutting off one or more side chains while keeping the side chain neighbors nearest the flexible ring;

perturbing every flexible corner of the isolated ring relative to an average ring plane;

reconnecting the chains cut off in the first step;

calculating torsional ranges for each rotatable bond using the differential distance equation;

responsive to the identification of points within the calculated torsional ranges, uniformly sampling the identified points; and

storing the sampled conformations.

2. The method of claim 1, further comprising the steps of, responsive to the calculated torsional range failing to meet the van der Waals distance constraints:

regrouping atoms of the compound;

calculating torsional ranges for each rotatable bond; and

responsive to a rotatable bond of the regrouped atoms meeting the van der Waals distance constraints, storing the conformation.

3. The method of claim 1, further comprising prior to calculating the torsional ranges, the additional step of removing compounds that have atom pairs whose minimum accessible distances are less than their van der Waal distance constraints.

4. The method of claim 1, wherein the conformational search is implemented in the Cartesian coordinate system.

5. The method of claim 1, wherein the conformational search is implemented in the internal coordinate system.

6. The method of claim 1, wherein the step of perturbing further comprises the step of applying several steps of energy minimization to optimize the ring structure.

7. The method of claim 6, wherein the step of reconnecting the side chains further comprises the step of applying several steps of energy minimization, wherein the optimization serves to optimize the structure of the compound and releases any bad van der Waals contacts between the side chains and rings.

8. A method for automating the identification of pharmacaphores from a large heterogeneous data set, the method comprising the steps of:

performing a conformational search for each compound;

storing all conformations generated during the conformational search;

searching for a most significant two-point descriptor using a t-test to analyze the two-point descriptor space;

responsive to finding the most significant two point descriptor:

storing two chemical features and distance between said features of the most significant two point descriptor, and

dividing all conformations into subsets according to the presence or absence of the stored two point descriptor;

searching all saved conformations of the compounds for which said the stored two point descriptor is present using the Ullman algorithm;

responsive to the presence of one or more conformations that satisfy the search in the prior step, performing a second conformational search on said conformations wherein the stored distance for the most significant two-point descriptor is included as an additional distance constraint;

responsive to one or more three-point descriptors generated during second conformational search, searching said three-point descriptors for a third most significant pharmacophoric point; and

performing a third conformational search on conformations which did not contain the most significant two point descriptor in order to locate a next most significant two point descriptor.

9. The method of claim 8, wherein the method steps are repeated until no more significant pharmacaphoric points has been attained.

10. The method of claim 8, wherein the method steps are repeated until a default number of significant pharmacaphoric points has been attained.

11. The method of claim 10, wherein the default number of significant points is five.

12. The method of claim 8, wherein the steps of storing includes storing conformations includes storing conformations using sparse matrix techniques.

13. The method of claim 8, wherein the step of storing includes storing data using dynamic memory management techniques.

14. The method of claim 8, wherein the steps of performing a conformational search comprises performing the conformational search of claim 1.

15. The method of claim 8, wherein the steps of performing a conformational search comprises performing the conformational search of claim 7.

16. A system for automating the identification of pharmacaphores from a large heterogeneous data set, the system comprising:

means for performing a conformational search for each compound in the data set;

coupled to the means for performing, means for storing all conformations generated during the conformational search; and

coupled to the means for storing, means for searching for the most significant descriptors in the descriptor space.

17. The system of claim 16, wherein the means for performing the conformation search includes:

a computer processor unit (CPU); and

a magnetic memory device having stored instructions which enable execution of the method of claim 1.

18. The system of claim 16, wherein the means for searching for the most significant descriptors includes means for performing a t-test.

19. A computer-readable medium containing a computer program for automating the identification of pharmacaphores from a large heterogeneous data set, said program containing instructions for directing the computer to execute the steps of:

performing a conformational search for each compound;

storing all conformations generated during the conformational search;

searching for a most significant two-point descriptor using a t-test to analyze the two-point descriptor space;

responsive to finding the most significant two point descriptor:

storing two chemical features and distance between said features of the most significant two point descriptor, and

dividing all conformations into subsets according to the presence or absence of the stored two point descriptor;

searching all saved conformations of the compounds for which said the stored two point descriptor is present using the Ullman algorithm;

responsive to the presence of one or more conformations that satisfy the search in the prior step, performing a second conformational search on said conformations wherein the stored distance for the most significant two-point descriptor is included as an additional distance constraint;

responsive to one or more three-point descriptors generated during second conformational search, searching said three-point descriptors for a third most significant pharmacophoric point; and

performing a third conformational search on conformations which did not contain the most significant two point descriptor in order to locate a next most significant two point descriptor.

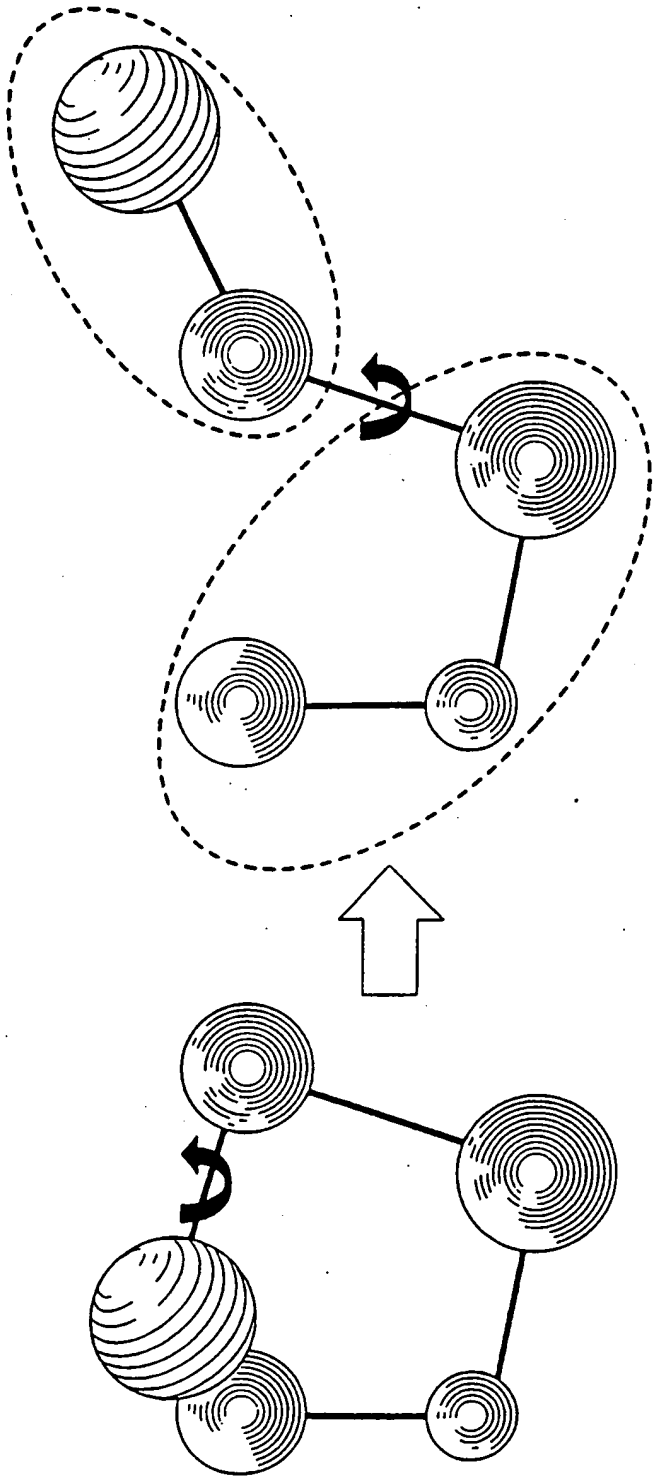


FIG. 1.

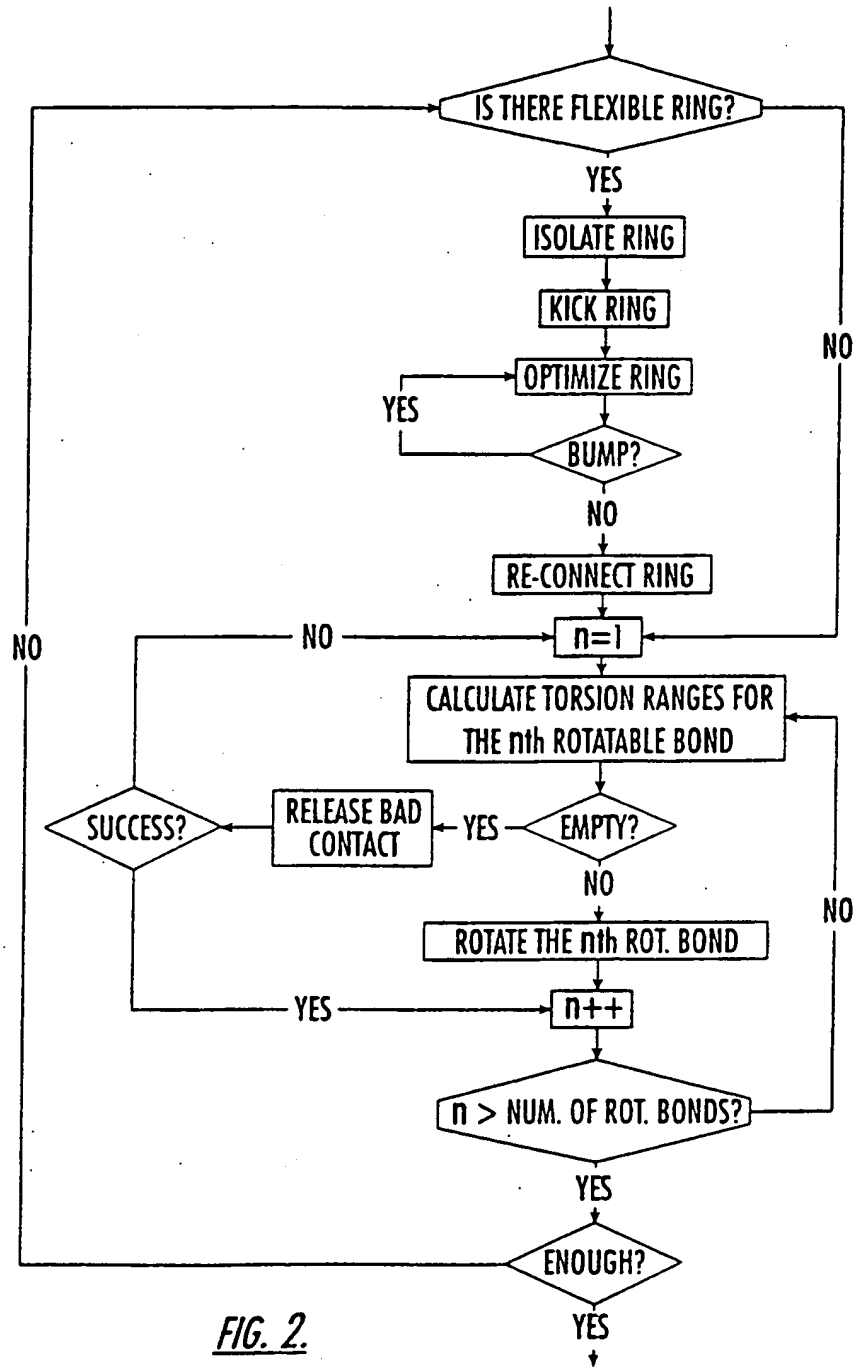

IS THERE FLEXIBLE RING?
YES
NO
ISOLATE RING
KICK RING
OPTIMIZE RING
YES
BUMP?
NO
RE-CONNECT RING
NO
n=1
NO
CALCULATE TORSION RANGES FOR THE nth ROTATABLE BOND
SUCCESS?
RELEASE BAD CONTACT
YES
EMPTY?
NO
ROTATE THE nth ROT. BOND
YES
n++
n > NUM. OF ROT. BONDS?
NO
YES
ENOUGH?
YES

FIG. 2.

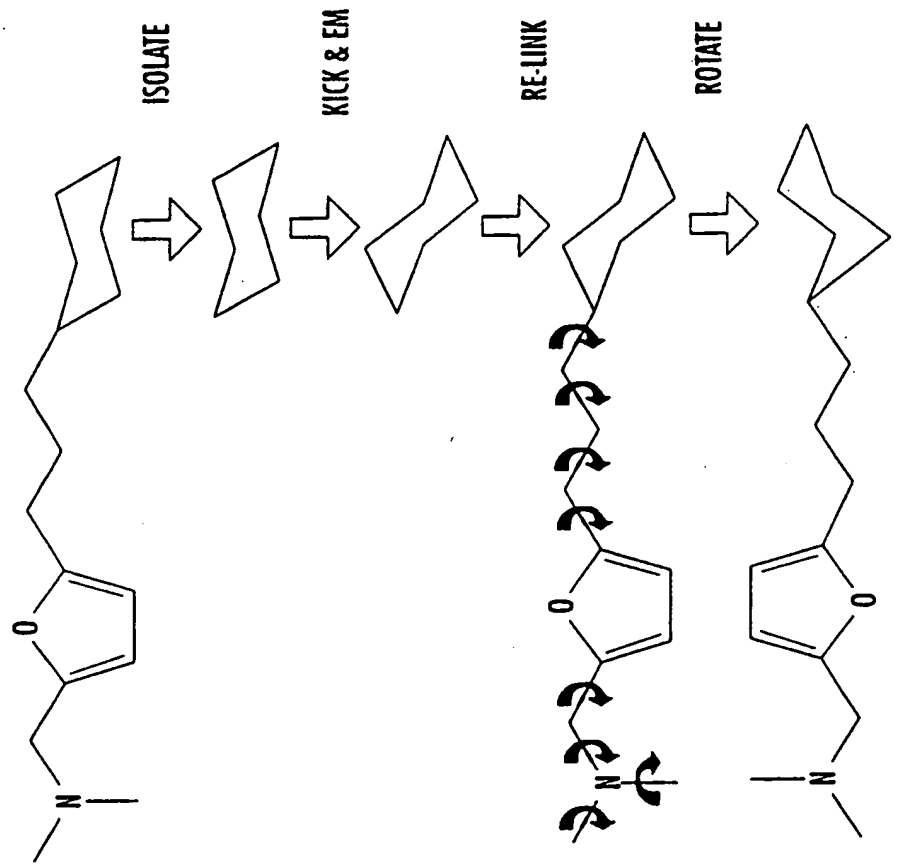


FIG. 3.

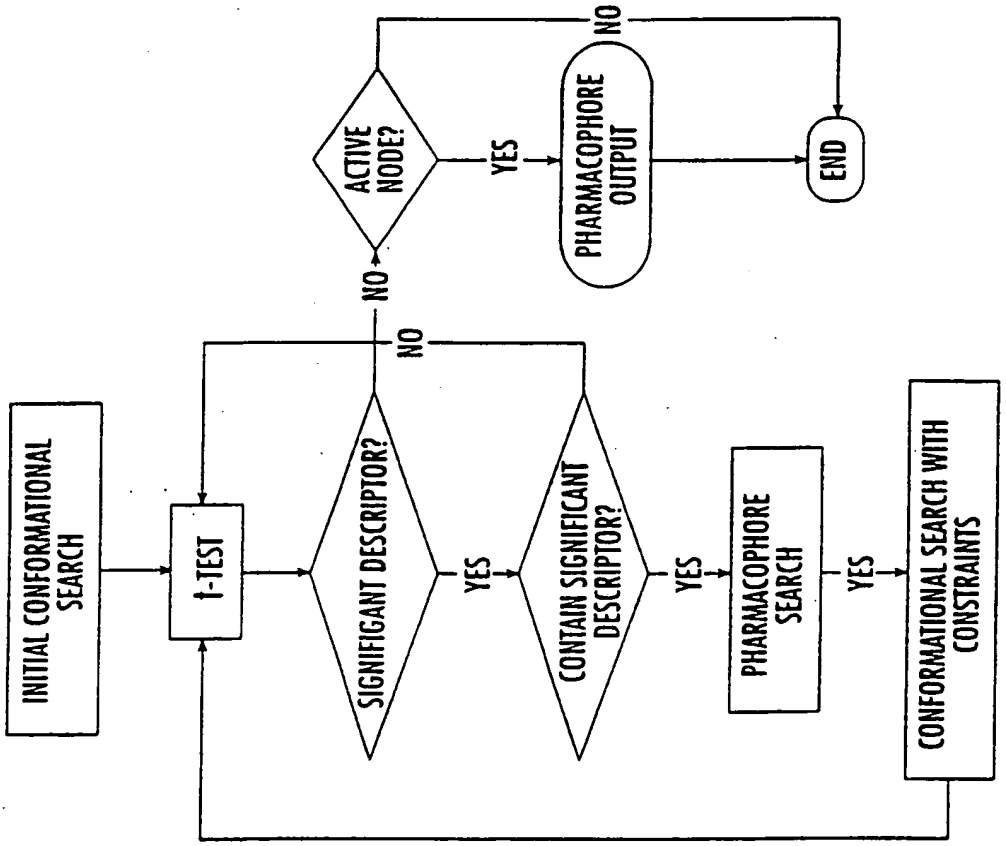

INITIAL CONFORMATIONAL SEARCH
t-TEST
SIGNIFICANT DESCRIPTOR?
CONTAIN SIGNIFICANT DESCRIPTOR?
PHARMACOPHORE SEARCH
CONFORMATIONAL SEARCH WITH CONSTRAINTS
ACTIVE NODE?
PHARMACOPHORE OUTPUT
END
YES
NO

FIG. 4.

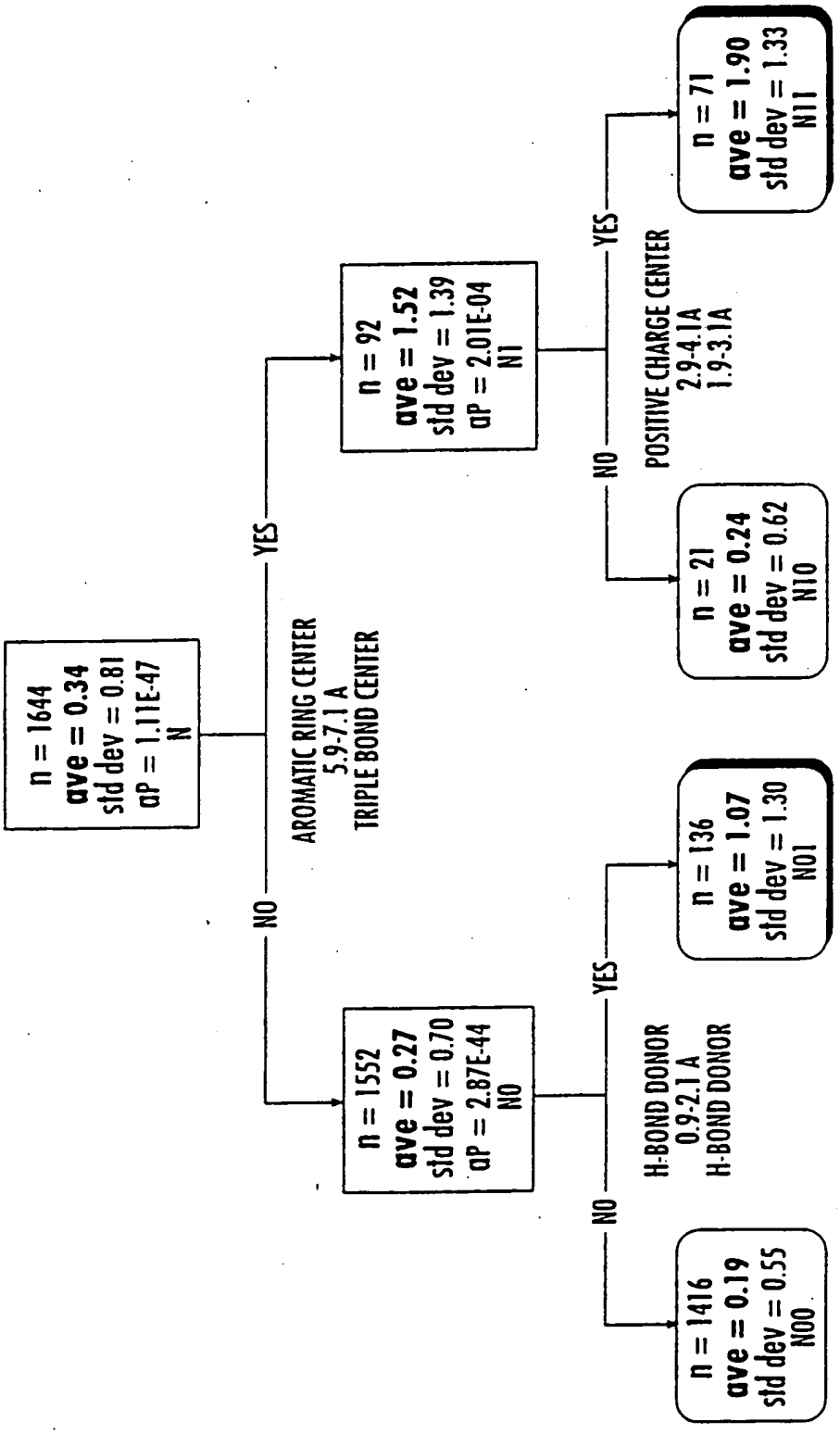
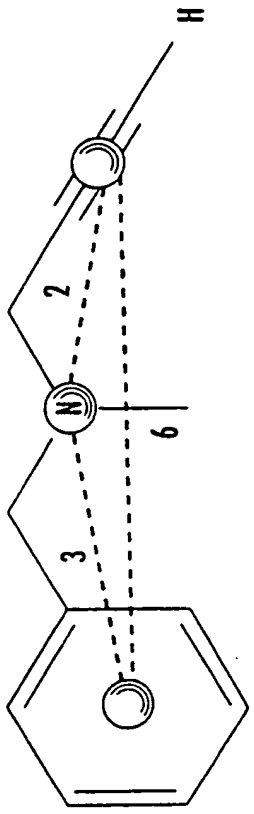
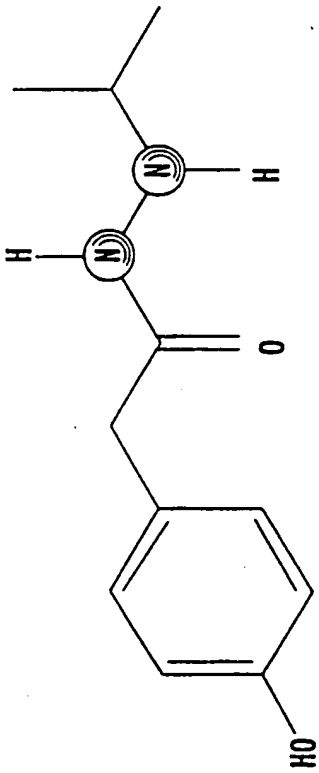

n = 1644
ave = 0.34
std dev = 0.81
αP = 1.11E-47
N
AROMATIC RING CENTER
5.9-7.1 A
TRIPLE BOND CENTER
YES
NO
n = 92
ave = 1.52
std dev = 1.39
αP = 2.01E-04
N1
n = 1552
ave = 0.27
std dev = 0.70
αP = 2.87E-44
N0
POSITIVE CHARGE CENTER
2.9-4.1A
1.9-3.1A
YES
NO
n = 71
ave = 1.90
std dev = 1.33
N11
n = 21
ave = 0.24
std dev = 0.62
N10
H-BOND DONOR
0.9-2.1 A
H-BOND DONOR
YES
NO
n = 136
ave = 1.07
std dev = 1.30
N01
n = 1416
ave = 0.19
std dev = 0.55
N00

FIG. 5.



AL19120 (IN NODE N11)



AL16432 (IN NODE N01)

FIG. 6.

TO FIG. 7B. 7 / 10

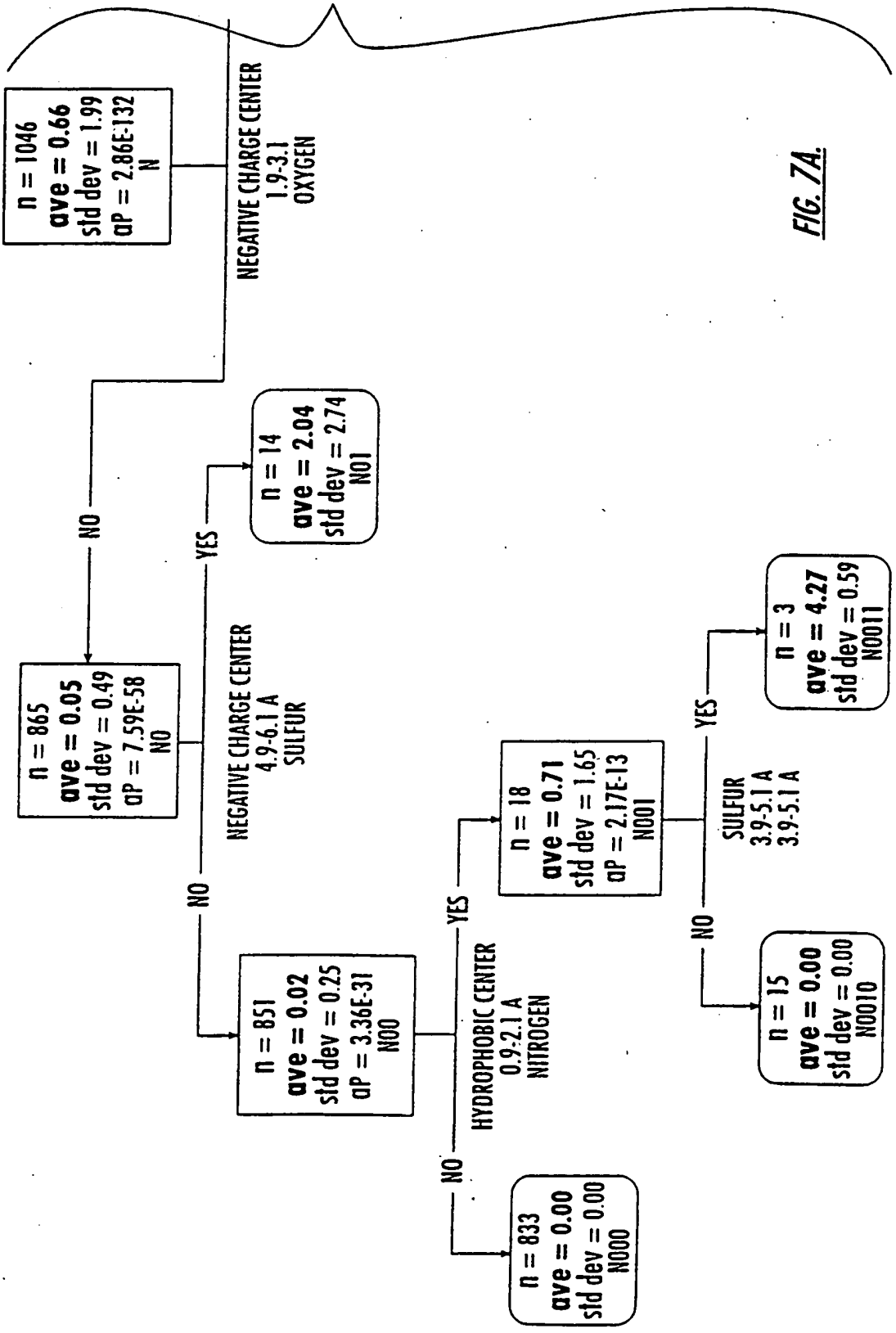


FIG. 7A.

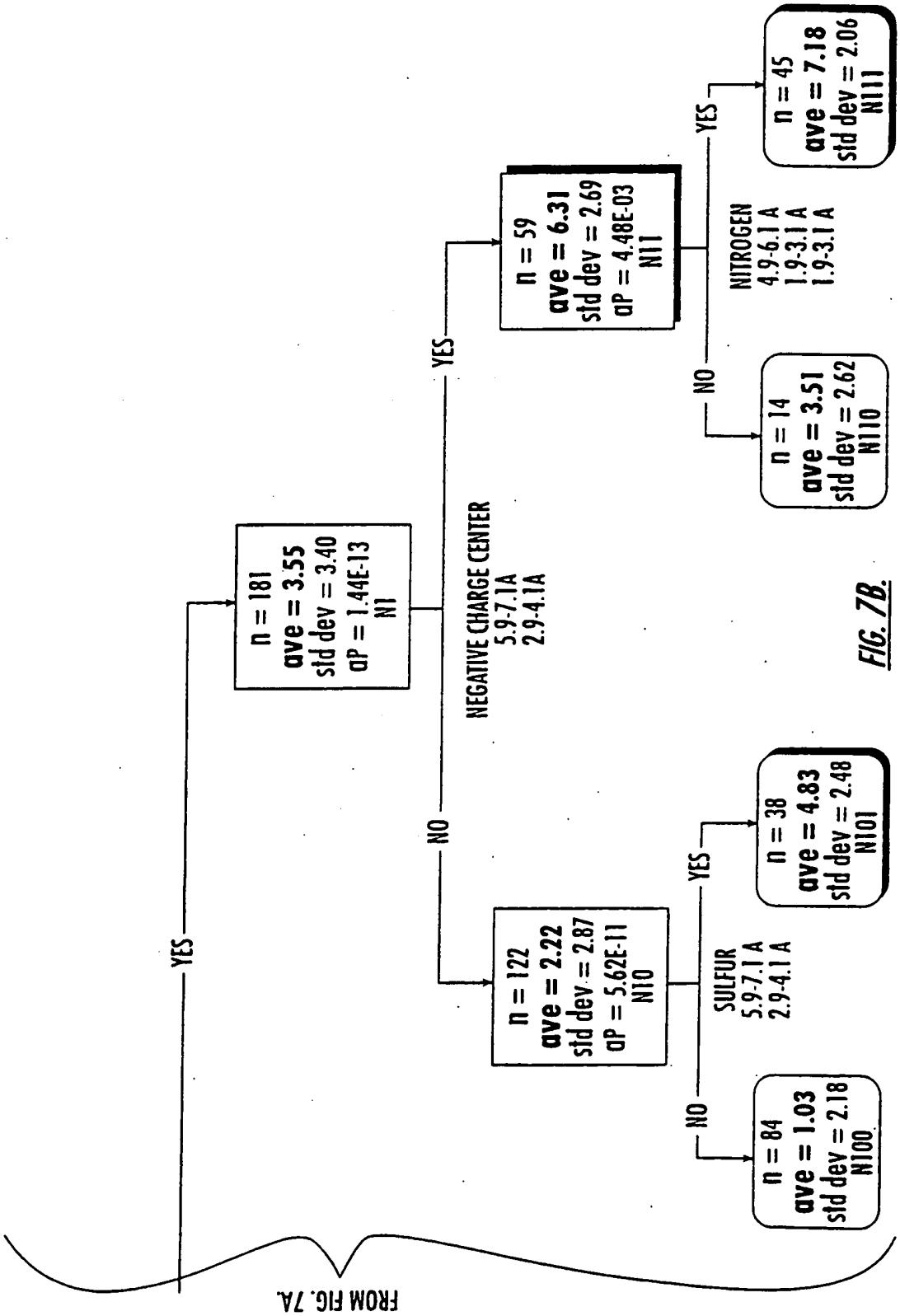

FROM FIG. 7A.
YES
n = 181
ave = 3.55
std dev = 3.40
aP = 1.44E-13
N1
NO
NEGATIVE CHARGE CENTER
5.9-7.1A
2.9-4.1A
YES
n = 59
ave = 6.31
std dev = 2.69
aP = 4.48E-03
N11
YES
n = 45
ave = 7.18
std dev = 2.06
N111
NO
NITROGEN
4.9-6.1 A
1.9-3.1 A
1.9-3.1 A
n = 14
ave = 3.51
std dev = 2.62
N110
NO
n = 122
ave = 2.22
std dev = 2.87
aP = 5.62E-11
N10
YES
n = 38
ave = 4.83
std dev = 2.48
N101
NO
SULFUR
5.9-7.1 A
2.9-4.1 A
n = 84
ave = 1.03
std dev = 2.18
N100

FIG. 7B.

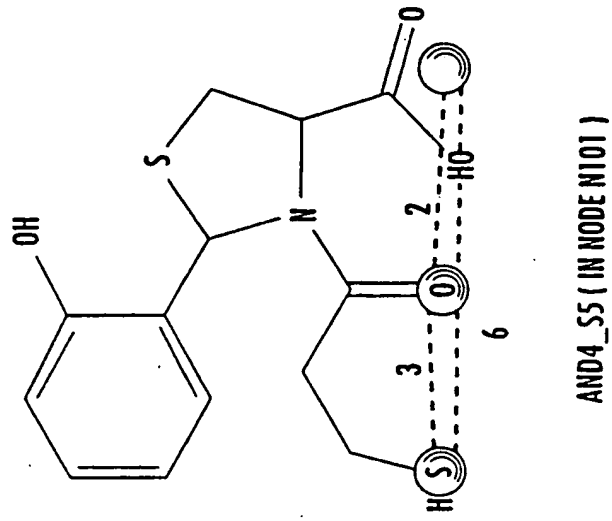


FIG. 8.

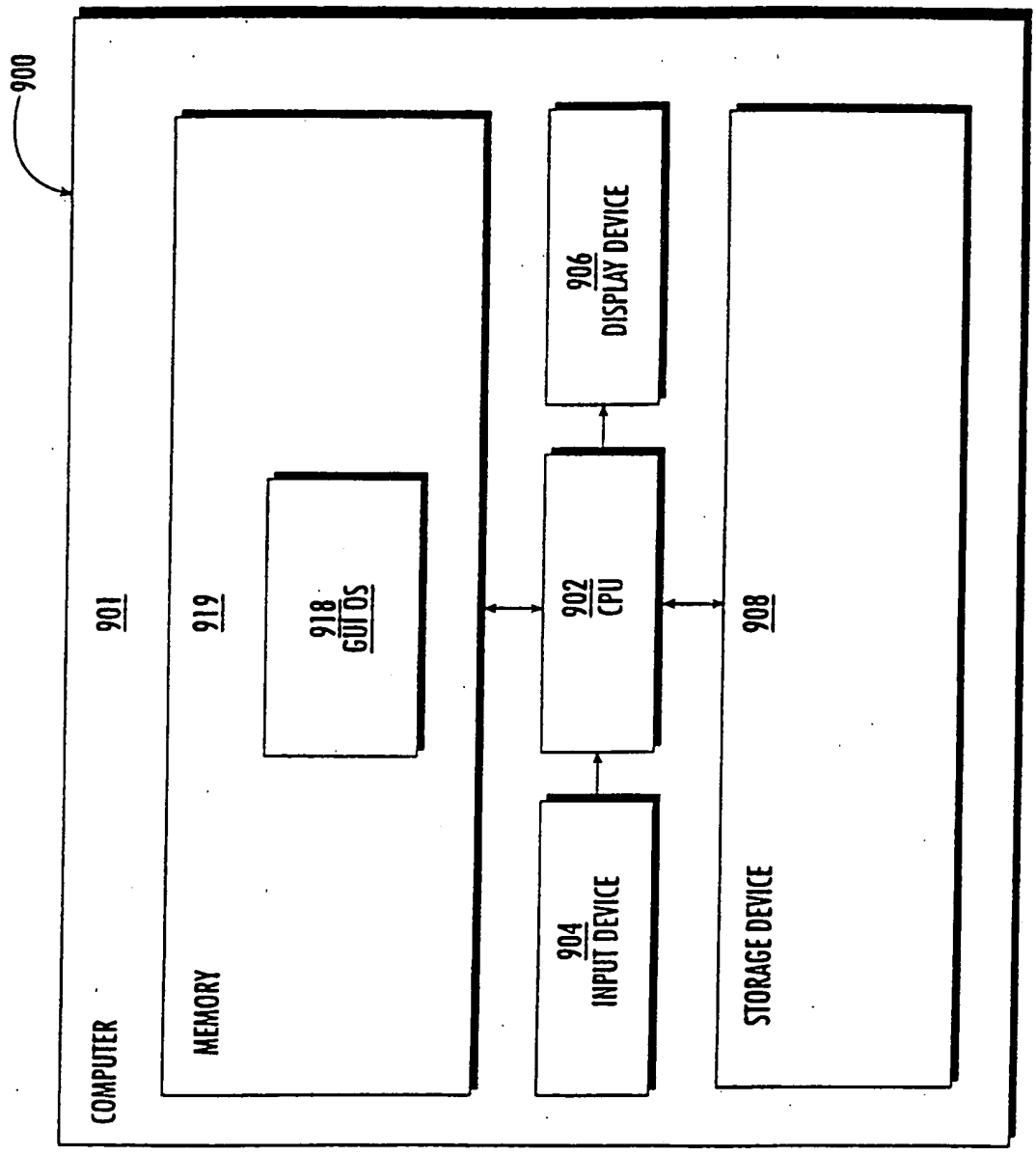

900
COMPUTER
901
MEMORY
919
918
GUI OS
906
DISPLAY DEVICE
902
CPU
904
INPUT DEVICE
908
STORAGE DEVICE

FIG. 9.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/25922

A. CLASSIFICATION OF SUBJECT MATTER
IPC(7) :G06F 15/00
US CL :364/413.01, 413.15, 413.16, 413.18, 488, 499
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 364/413.01, 413.15, 413.16, 413.18, 488, 499

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

STN, CAPLUS, USPATFULL
search terms: generic match algorithm, CHARMM, computer modeling, conformational search.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	MACKERELL, JR. et al. All-Atom Empirical Potential for Modeling and Dynamics Studies of Proteins. J. Phys. Chem. B. 1998, Vol. 102, No. 18, pages 3586-3616, see entire document	1-7, 16-19
Y	XU, J. GMA: A Generic Match Algorithm for Structural Homomorphism, Isomorphism, and Maximal Common Substructure Match and Its Applications. J. Chem. Inf. Comput. Sci. 1996, Vol. 36, No. 1, pages 25-34, see entire document.	1-7
Y	BARNUM et al. Identification of Common Functional Configurations Among Molecules. J. Chem. Inf. Comput. Sci. 1996, Vol. 36, pages 563-571, see entire document.	8-19

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search	Date of mailing of the international search report
21 JANUARY 2000	08 FEB 2000

Name and mailing address of the ISA/US	Authorized officer
Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231	MICHAEL WOODWARD
Facsimile No. (703) 305-3230	Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*